

Properties of the EDF

- Values between any two consecutive samples, x_k and x_{k+1} cannot be simulated, nor can values smaller than the sample minimum, x_1 , or larger than the sample maximum, x_n , be generated, i.e., $x \geq x_1$ and $x \leq x_n$.
- The mean of the EDF is equal to the sample mean. The variance of the EDF mean is always smaller than the variance of the sample mean; it is equal to $(n-1)/n$ times the variance of the sample mean.
- The variance of the EDF is equal to $(n-1)/n$ times the sample variance.
- Expected values of simulated EDF percentiles are equal to the sample percentiles.
- If the underlying distribution is skewed to the right, the EDF will tend to under-estimate the true mean and variance.

Variations of the EDF

Linearized EDF

linearly extrapolating between observations

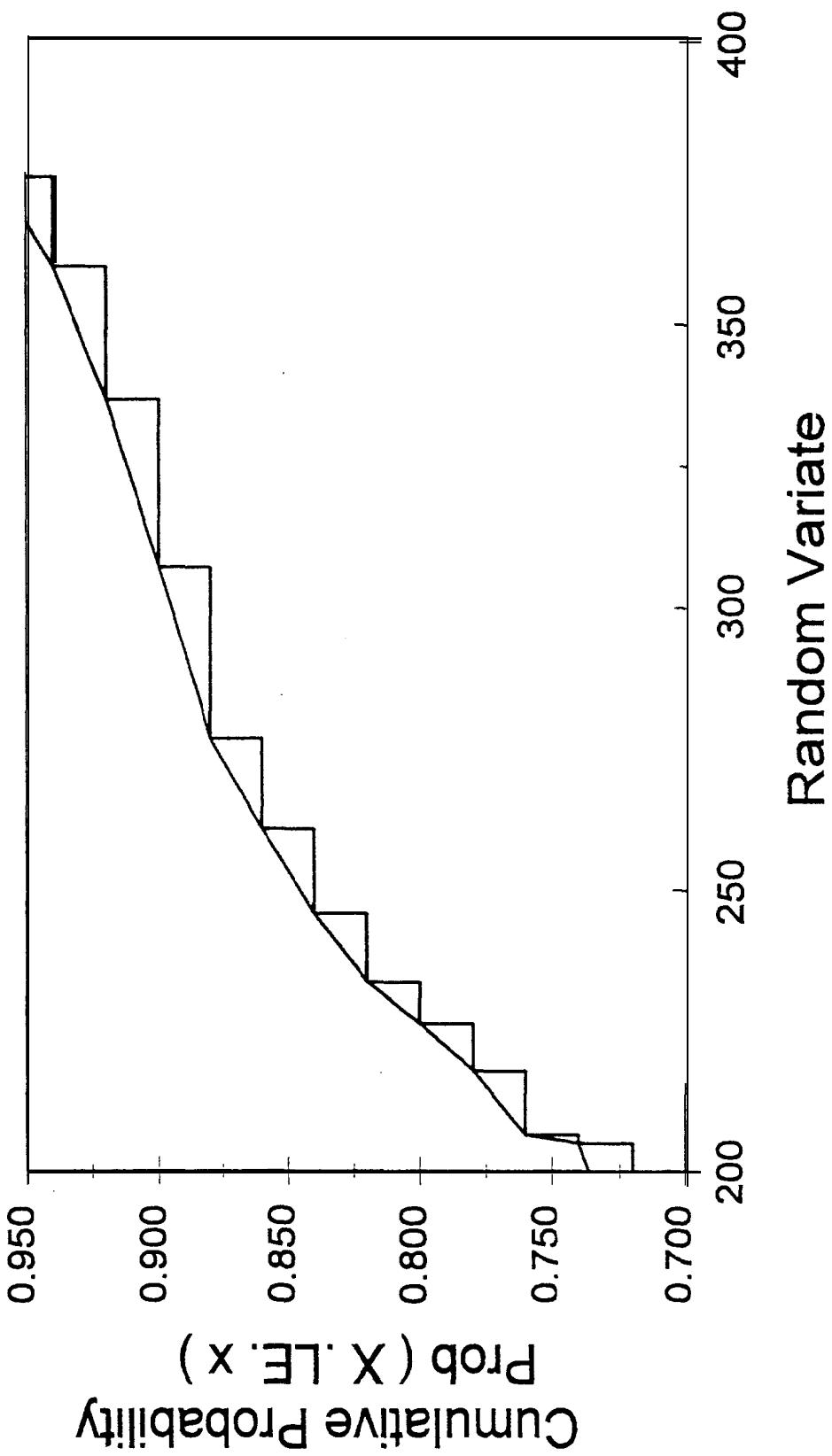
Extended EDF

based on expert judgement, adding lower & upper tails to the data to reflect “**a more realistic range**” of the exposure variable (EDFs produce tails that are too short)

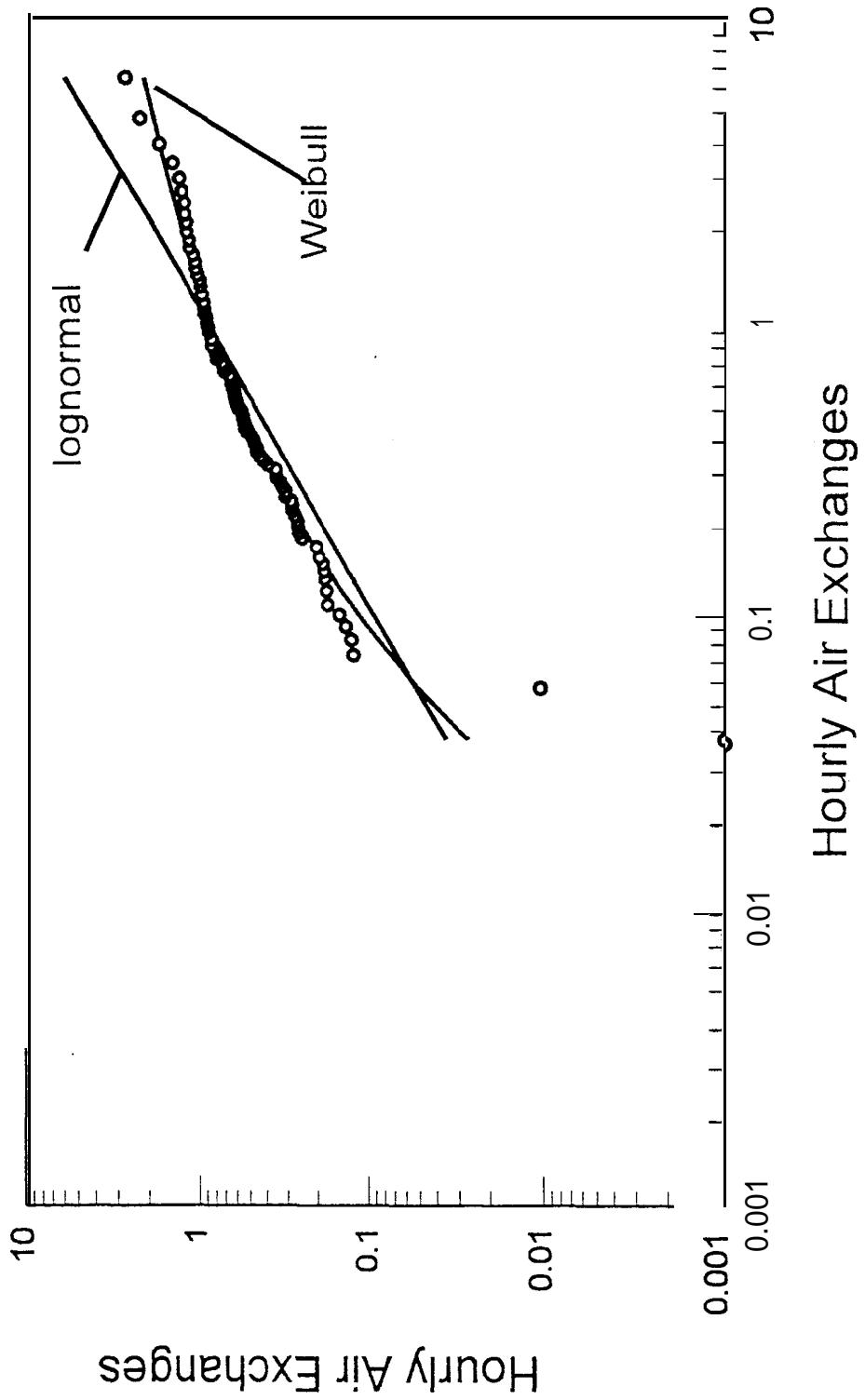
Mixed Exponential

based on extreme value theory, adding an exponential upper tail to the EDF to model the exponential behavior of many continuous, unbounded distributions

Figure 4. Comparison of Basic EDF and Linearly Interpolated EDF



Comparison of Fitted Lognormal and Weibull Distributions to ACH Data



| Statistic | ACH Sample N = 90 | Best Fit | |
|-----------|-------------------------|-------------------|----------------|
| | | Linearized EDF | Weibull PDF |
| mean | 0.6822 | 0.6821 | 0.6747 |
| variance | 0.2387 | 0.2358 | 0.2089 |
| skewness | 1.4638 | 1.4890 | 1.2426 |
| kurtosis | 6.6290 | 6.7845 | 5.6966 |
| 5% | 0.1334 | 0.1320 | 0.1307 |
| 10% | 0.1839 | 0.1840 | 0.1840 |
| 50% | 0.6020 | 0.6160 | 0.6032 |
| 90% | 1.2423 | 1.2390 | 1.2398 |
| 95% | 1.3556 | 1.3820 | 1.3600 |

EDF Questions

- Are there circumstances in which an EDF is preferred over a TDF?
- Are there situations in which an EDF should not be used?
- When an EDF is used, should the linearized, extended or mixed versions be used?

Goodness of Fit Questions

Generally, we should pick the simplest analytic distribution not rejected by the data..... **But**, rejection depends on the statistic chosen and an arbitrary level of statistical significance.

- What role should the GoF statistic and its p-value (when it is available) play in that judgment?
- What role should graphical assessments of fit play?
- When none of the standard distributions fit well, should we investigate more flexible families of distributions, e.g. four parameter gamma, four parameter F, mixtures, etc.?

OBJECTIVES

- Illustrate use of *Bayesian statistical methods*
 - variability in an exposure factor (arsenic concentration) is represented by a probability distribution model
 - uncertainty is characterized by the probability distribution function of the model parameters
- Illustrate use of probability distribution model with *covariates* (*explanatory variables*)
 - allowing extrapolation to different target populations